LABTHOUGHTS: A VIDEO DATASET TOWARD UNDERSTANDING GOAL-DIRECTED COGNITION IN THE LAB

Anonymous authors
Paper under double-blind review

ABSTRACT

In scientific research, understanding experiments in the laboratory calls for not only physical scene understanding, but also for inferring how actions relate to a scientist's goals, assessments, and evolving plans. We introduce LabThoughts, a dataset of naturalistic egocentric video from in-situ wet-lab research in organic chemistry, biochemistry, and microfluidics. LabThoughts also includes concurrent think-aloud narration during experiments and post-experiment interviews, providing unusually direct evidence about the reasoning and intentions underlying behavior in the laboratory. We annotate these transcripts to highlight statements related to different components of **goal-directed cognition**; specifically **assessments** of the experimental state, **rationales** behind decisions, and **intentions**. Using these annotated transcripts, we construct a video question-answering evaluation in which questions are grounded in the expressed cognitive states. Initial experiments with multimodal language models show substantially different performance across question types and context conditions: concrete intentions are generally easier than questions about rationales and assessments, while providing explicit action descriptions improves performance. These results suggest that LabThoughts provides a useful setting for studying how well video models connect observed laboratory activity to the underlying goal-directed cognition.

1 INTRODUCTION

Recent advances in multimodal large language models' (MLLMs) abilities in processing, recognizing, and reasoning about activity in video have opened the possibility of assistive systems that can understand and assist with users' real-world activities, enabled by devices like camera-equipped wearables. However, providing actually effective assistance often requires understanding beyond what is directly observable in the physical scene. AI systems must understand a person's goals and intentions in order to provide meaningful assistance, a task which is very challenging for even large, modern models (Veerabadran et al., 2025). This is especially challenging in the context of *scientific laboratories, where AI systems must combine complex domain knowledge with video understanding, and where understanding a scientist's actions requires reasoning about both the scientist's current goal and their current understanding of the experimental state*. We refer to these thought processes as **scientific goal-directed cognition**, i.e., cognitive processes organized around the pursuit of the scientific goal. An assistive system that recognizes laboratory actions but lacks this broader goal-directed context may have limited ability to support activities such as training new lab members or troubleshooting unexpected results and reproducibility failures.

However, such cognitive states are difficult to observe directly. To support the study of goal-directed cognition, we introduce LabThoughts, a dataset designed to provide evidence about goal-directed cognition through both concurrent think-aloud narration and post-experiment interviews, collected while researchers perform their own ongoing laboratory work. LabThoughts consists of long-form egocentric video, averaging 63 minutes per session, recorded during naturally occurring wet lab research. The dataset contains five experiments conducted by four Ph.D. researchers across chemistry, biochemistry, and microfluidics. Unlike datasets centered on prescribed actions or fixed protocols, LabThoughts captures researchers pursuing their existing research goals in situ. Participants narrate their thoughts while working and later complete interviews reflecting on their decisions and experimental progress. We annotate these transcripts for three forms of goal-directed cognition: **assessment**, **rationale** (reasoning), and **intent** (see Figure 2). This annotation framework is inspired

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

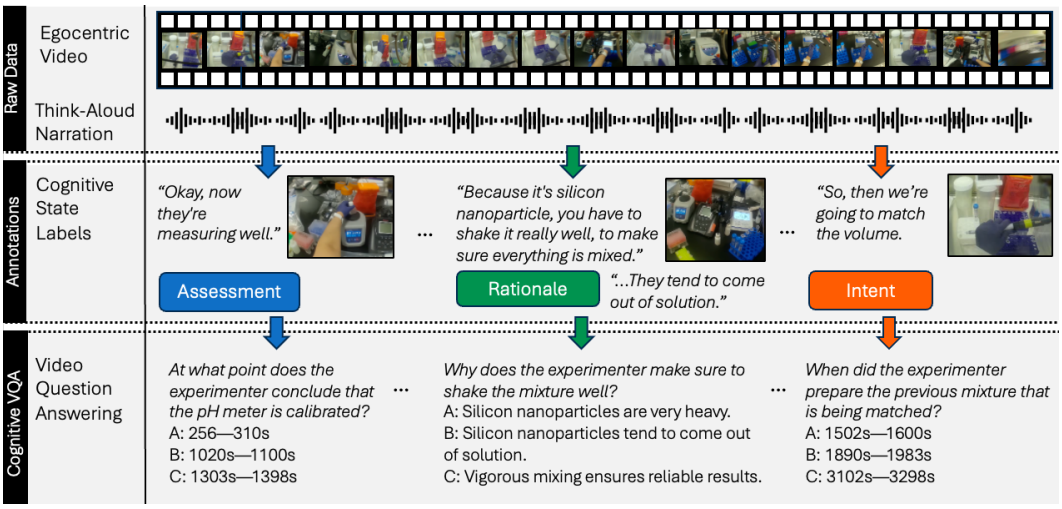


Figure 1: **Overview of LabThoughts.** Our VQA evaluation draws its questions and answers from the thoughts expressed by the experimenter. These questions have a **cognitive dependency**: answering requires interpreting the experimenter's cognitive state as well as video understanding.

by classic cognitive psychology accounts of problem-solving in which humans assess the current state, identify intermediate subgoals, and select actions that move them toward a desired goal state (Simon & Newell, 1971).

From the LabThoughts dataset, we construct a cognitively focused video question-answering (VQA) evaluation, consisting of a set of over 500 multiple-choice questions. Each question is grounded in an annotated statement in which the experimenter explicitly expresses a specific instance of goal-directed cognition. We refer to this question construction approach as *cognitive dependency*: the target answer is grounded in information expressed in the experimenter's cognitive-state annotation. During evaluation, models do not receive the privileged think-aloud transcript, allowing us to test whether they can infer this information from the video and other available context (see Figure 1 for examples). We create both questions that directly inquire about the experimenter's cognition, and questions that require temporal localization. We select temporal localization segments based on timestamps where the experimenter explicitly expresses an intent or assessment. As a result, the temporal questions are designed to localize events that are meaningful with respect to the experimenter's underlying goal-directed cognition, rather than arbitrary actions or state changes.

Initial evaluations on the VQA set show that accuracy varies by both cognitive-state category and available context. Our results suggest that questions grounded in expressions of concrete intentions are generally easier than those grounded in rationales or assessments of experiment state. We also find that providing explicit textual descriptions of actions results in accuracy improvement, suggesting that failures to recognize or integrate activity may contribute to difficulty with higher-level cognitively focused questions.

Our contributions are: (1) LabThoughts, a naturalistic multimodal dataset designed to support the study of goal-directed cognition during authentic scientific research activity; (2) **Annotation framework** for labeling goal-directed cognition related statements in think-aloud narrations during laboratory work as **assessments**, **rationales**, and **intents**; (3) **Cognitively grounded VQA**, a question-generation framework that grounds VQA items in the annotated statements, preserving the answer's dependency on the cognitive process, and an accompanying initial evaluation of multimodal models.

2 Background

Recent interest in the potential for AI-assisted laboratory work has led to an effort to collect video data of wet lab activity. Many of these efforts focus on short actions or protocols (Gabrieli et al., 2024; Cong et al., 2025; Xu et al., 2026), with the goal of evaluating action and intent recognition in

Dataset	Clip length	Total duration	VQA	Cognitive Commentary	Domain(s)	Activity type
Lab Actions (Gabrieli et al., 2024)	4 s	∼7 h	✗	✗	Chemistry	Isolated protocols
EgoExoLearn (Three lab tasks) (Huang et al., 2024)	7–22 min	∼29 h†	1,454	▲ (action; intent only)	Chemistry, medicine, biology	Isolated protocols
LabSuperVision (Cong et al., 2025)	2–10 min	∼16 h†	✗	✗	Biology	Isolated protocols
ExpVid (Xu et al., 2026)	8 min	∼57 h†	7,800	✗	Various	Isolated protocols (from JOVE database)
LabThoughts (Ours)	**38–88 min**	∼5 h	574	✓	Chemistry, medicine, microfluidics	**Naturally occurring research activity**

Table 1: **Comparison with laboratory video datasets.** LabThoughts provides long-form videos with concurrent think-aloud commentary. Unlike prior datasets based on prescribed actions, demonstration following, prescribed procedures, or published protocols, LabThoughts captures open-ended research activity. †Approximate duration estimated based on reported video average length and total video count.

the laboratory. LabThoughts is unique in the length of the videos – experiments are all at least 38 minutes long and up to 88 minutes long – and the commentary provided by experimenters. It is also unique in capturing naturally occurring research activity, as opposed to isolated given protocols. See Table 1 for a summary.

Goal-Directed Cognition. The study of the cognitive processes in humans that enable the pursuit of goals largely originate with Simon & Newell (1971). Their theoretical framework of goal-oriented cognition, both inspired by and validated with human think-aloud data, centers on states (e.g., desired goal states) and operators (e.g., actions) that transition between these states. Goal-oriented cognition is the search for and selection of desired states and the best operators to achieve those states. Cognitive architectures designed to model human cognition, such as Soar (Laird, 2012), have incorporated and extended this framework. In the decades since Simon and Newell's work, the study of goal-directed cognition and problem-solving in humans has increasingly been integrated with other cognitive constructs like executive functioning (Diamond, 2013), rather than developed further as a separate cognitive module. We therefore focus on Simon and Newell's framework, as it is an explicit and relatively self-contained account of goal-directed cognition.

Figure 2 depicts our operationalization of this goal-directed cognition loosely organized into three states: goal-state **assessments**; **reasoning** and planning; and **intent** formation, each of which moves a person closer to taking an action intended to advance their goals. While these states are not typically observable, our approach approximates them through narrations of thoughts during the experiment, using a think-aloud protocol in which experimenters narrate their thoughts (Ericsson & Simon, 1993).

Video Understanding. In-the-wild video naturally contextualizes a person's actions within their environment. Significant progress in human understanding has been enabled through large-scale video collection efforts of both unstructured capture (Soomro et al., 2012; Carreira & Zisserman, 2017; Miech et al., 2019; Grauman et al., 2022; Wang et al., 2024; Yang et al., 2025b) and prescribed tasks in natural environments (Damen et al., 2018; Grauman et al., 2024; Huang et al., 2024; Ma et al., 2025; Perrett et al., 2025; Xu et al., 2026). The visual stream, accompanying audio, and grounded narration present in these datasets are essential to training and enabling excellent video-reasoning capabilities of Multimodal Large Language Models (MLLMs) (Liu et al., 2023; Yang et al., 2025a; Google DeepMind, 2026a;b).

Visual Question Answering (VQA) (Antol et al., 2015) provides a way to evaluate MLLMs despite their nature of open-ended response, and for especially large models their direct inaccessibility due to



Figure 2: **Goal-directed cognition overview.** We conceptualize the goal-directed cognition behind action selection as follows. An **assessment** of the current state to determine progress toward a goal or subgoal and inform decision-making; **reasoning** about the current state and the goal identifies the next desirable state and a plan to reach it (expressed as rationales); this leads to the formation of a concrete **intent** to pursue; and finally, an action to work toward that intent. Colors match the annotation labels in Figure 1. Goal-state assessment, reasoning and planning are especially challenging in scientific research settings, given the complexity of the goals.

Label	Definition	Example think-aloud statement
Assessment	Evaluates the current state relative to an active subgoal	*"That's the color we're looking for; that's a really good sign."*
Rationale	Explains the decision-making process and reasoning	*"I'll put that in a round bottom flask so that way I can rotovap it easily."*
Intent	Expresses the short-term subgoal behind future actions	*"So then I'm going to do one more kind of sodium carbonate wash."*

Figure 3: **Annotation labels.** From the experimenter's think-aloud narration, we label three observable components of the goal-directed cognition that enables the experimenter to satisfactorily complete the experiment. The same colors are used for each label throughout the paper.

API-based interaction. By carefully designing a multiple-choice exam, both the MLLM and prompt can be probed on targeted reasoning capabilities, such as what the person is doing, how it's being done, why the person is doing it (Perrett et al., 2025), and what they might do next (Li et al., 2025b). Best-practices for creating effective questions and answers has been a focus of VQA research, such as balancing visual priors with language priors (Goyal et al., 2017), or requiring answers to depend on understanding a long temporal sequence (Mangalam et al., 2023). Our approach builds upon these practices, requiring the answer to depend on the cognitive process of the video's subject.

3 LabThoughts Dataset

3.1 Data Collection

Egocentric laboratory videos were collected via either a headmounted camera[1], or via a smartphone attached to a cheststrap. The experimenter selected a preferred approach based on the laboratory environment, any required personal protective equipment, and personal comfort. Experimenters were instructed to speak their thoughts out loud while recording, following think-aloud protocol instructions (Ericsson & Simon, 1993), but to otherwise conduct their usual research. After the research session was completed, experimenters completed an hour-long follow-up interview to elaborate on their cognitive processes during the experiment. Participants were recruited through word-of-mouth and snowball sampling at our institution. Participants were compensated $150 for their participation. The project was approved by the IRB.

Transcript annotation. Think-aloud transcripts from the experiments are annotated by the authors according to the following schema (as outlined in Figure 3):

- **Assessment** Statements that judge the current experiment state; e.g., assess whether something has gone wrong or whether a reaction has begun. Examples: "That's the color we're looking for, that's a really good sign," "We're getting a lot of product in here now."
- **Rationale** Statements that expose the underlying reasoning or rationale behind a decision. Examples: "I'm done with this now so I don't need to put it back in the ice bath," "It's a small flask, so I got to choose a [stir bar] that looks small."

[1]The Ordro EP6 Plus; https://ordro.online/.

Experiment	Domain	Duration	Assess.	Rationale	Intent	Total	Label mix
Nitration	Chemistry	43:13	53	102	19	174	
Tin reduction	Chemistry	1:11:45	56	108	21	185	
Imine condensation	Chemistry	38:43	31	67	25	123	
Urine oligomer assay	Biochemistry	1:28:29	20	90	23	133	
Surfactant coalescence	Microfluidics	1:12:41	29	34	19	82	
Overall			**189**	**401**	**107**	**697**	
Share (%)			27.1	57.5	15.4	100	

Table 2: **Annotated transcript lines per experiment.** Number of think-aloud lines labeled assessment, rationale, or intent in each experiment; the bar shows each experiment's label mix. The majority of annotated statements are rationales.

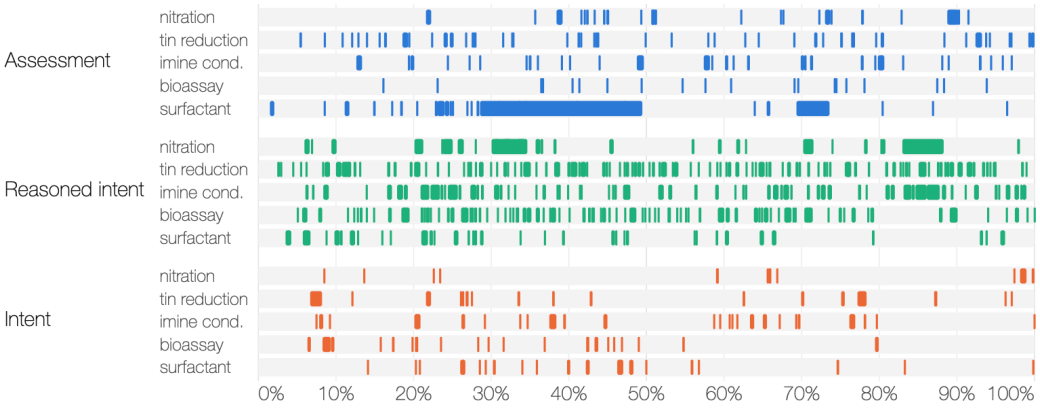


Figure 4: **Annotation timeline.** Position of each assessment, rationale, and intent statement over the course of each experiment. Rationales are expressed most frequently in the dataset; intents expressed without providing an associated rationale are relatively rare. Assessments occur regularly throughout each experiment.

- **Intent** Statements that express a subgoal and the intention to move toward that subgoal. Examples: "So I'm going to play with the stir rate a little bit. I think being in that 650 range is about where I want to be."

Every line of the transcripts was examined and assigned one of the above labels when applicable. These labels then serve to guide question and answer generation. Frequencies of all annotations, broken down by experiment, are given in Table 2; the temporal distribution of labels in each session is in Figure 4.

3.2 Experiments depicted in LabThoughts

Videos are sourced from five separate experiments conducted by four PhD students. Table 2 lists each experiment with its domain and annotation counts. The five experiments are described briefly below.

1. Nitration (chemistry). Nitrating a protected diamine while leaving the protected amines unreactive. The recording includes reagent preparation, controlled addition, filtration, and washing. It captures fine-grained experimental judgment such as monitoring reaction progress from color and dissolution.

2. Tin reduction (chemistry). Reducing nitro groups to amines. The session includes troubleshooting and perceptual judgment, including responding to an ineffective stir bar, responding to a clogged filter cake, and using solution appearance to assess e.g. dissolution and product quality.

3. Imine condensation (chemistry). Condensing reduced amine with a ketone partner to form a tritopic macromolecule. The recording shows air-free technique and procedural judgment, including Schlenk-line operation, solvent sparging and syringe transfer, and selective recovery of supernatant.

4. Urine oligomer assay (biochemistry). Testing sodium concentrations to identify conditions that maximize the Nano-QuIC signal for detecting trace protein aggregates in human urine. The recording involves contamination control, cognitive strategies such as rack position as a visual signal of progress, visual confirmation of microliter transfers, and recovery from an overshot titration.

5. Surfactant coalescence (microfluidics). Measuring how a surfactant–salt mixture affects bubble coalescence across different flow conditions in a microfluidic device. The recording covers solution and device preparation and syringe-pump control, with frequent real-time judgments of when bubble formation is stable, interpreting channel behavior, and adapting to backflow and bubble adhesion.

3.3 Cognitively grounded VQA construction

We construct a video question-answering evaluation from the goal-directed cognition expressed in the think-aloud transcripts. The construction pipeline begins with transcript statements annotated as **assessments**, **rationales**, or **intents** as detailed in §3.1. From these statements we build the question types shown in Figure 5.

Cognitive grounding. For each annotated statement, Gemini 3.1 Pro is prompted[2] to generate a question and correct answer grounded in both the video and the information expressed in that statement. Because the target answer is grounded in information explicitly expressed in an annotated statement of goal-directed cognition, we refer to this construction criterion as cognitive dependency. (During evaluation, the annotated statements are withheld.)

Distractor generation. For each question, we generate two incorrect answers, again using Gemini 3.1 Pro.[3] We use an adversarial generation procedure to produce plausible and tempting distractors. One distractor is generated without access to the experiment context and is instructed to provide a plausible but incorrect answer grounded in general scientific knowledge, while the second is generated with access to the video and is instructed to provide a plausible but incorrect answer grounded in the observed activity. The result is multiple choice questions with three possible answers, consisting of one transcript-grounded target answer and two distractors.

Context-free filtering. To reduce questions that can be answered or correctly guessed using general scientific knowledge rather than information specific to the experiment, we apply a context-free filtering step. Specifically, we present Gemini 3.1 Pro with each question and no experiment-specific context three times. We remove questions that are answered correctly in at least two of the three runs. This filtering step is intended to remove questions that are trivially answerable either through scientific knowledge or cues in the structure of the answers.

Human quality check. To estimate the quality of the generated questions, an author in the sciences assessed a randomly selected subset of 90 questions. 15 of the questions (16.7%) were found to have a distractor that was not clearly worse than the correct answer. The presence of these correct or nearly correct distractors may limit the ability of the VQA to distinguish between hypothetical systems with very high accuracy (in practice, today's MLLMs are well below this limit, see §5).

When evaluating a model on the constructed questions, models do not have access to the privileged think-aloud statements that form the basis of the VQA generation pipeline. Models must instead infer the target answer based on the video itself and any additional context supplied in the experimental condition. While the goal-directed cognition questions directly inquire about the experimenter's cognitive state, the pipeline also ensures that temporal localization questions specifically target segments annotated with the experimenter's statements related to goal-directed cognition. This ensures that answering the temporal localization questions requires understanding preparation and recognition of states relevant to the experimental goal, rather than standard activity recognition.

4 Experimental Setup

Models. We use Gemini 3.1 Pro (Google DeepMind, 2026b) and Gemini 3.1 Flash Lite (Google DeepMind, 2026a). Gemini models are accessed via LiteLLM[4] and the VertexAI API using default

[2] Prompts can be found in Appendix C.1.
[3] Prompts can be found in Appendix C.2.
[4] https://github.com/BerriAI/litellm

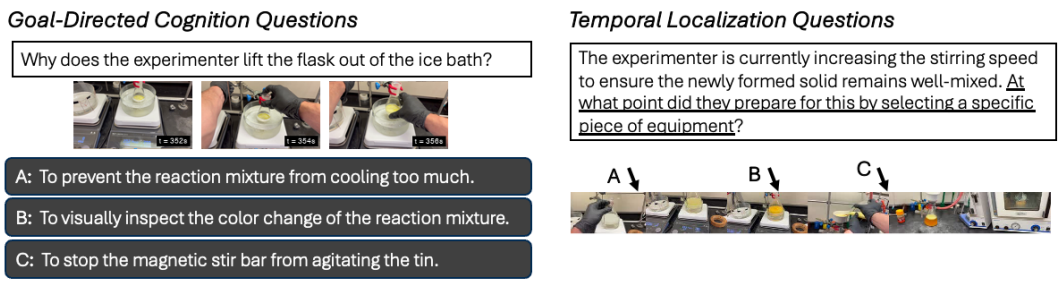


Figure 5: **Question types.** The first type (left) explicitly assesses scientific goal-directed cognition. The second (right) implicitly assesses understanding of the scientist's internal plans and subgoals, through specific subgoal states and previous actions that enabled a given action.

generation settings.[5] We also use one open-source model, Qwen3-VL-32B-Instruct (Yang et al., 2025a), with FP8 quantization, served by vLLM (Kwon et al., 2023) on two NVIDIA RTX A6000 GPUs (48 GB). We resize frames to 480px width. Gemini models take video inputs; we give Gemini the videos compressed to 0.5fps and 32 CRF. For Qwen, we extract frames as follows: Frames are sampled at 0.2 fps, the highest rate that allows us to fit the full-length videos fit within the context window, and each frame is paired with a text timestamp.

Evaluation settings. We evaluate models without access to the privileged think-aloud statements used to construct the VQA. We specifically evaluate three context conditions:

- **Video only.** The model receives the experiment video and question. This setting is motivated by providing the minimum context that a camera-enabled assistive system can expect.
- **Video + goal.** In addition to the video, the model receives the experimenter's description of the experimental goal. This setting is motivated by previous findings that MLLMs have difficulty inferring goals from observation alone (Veerabadran et al., 2025; Li et al., 2025a).
- **Video + goal + actions.** The model additionally receives a textual description of the actions performed during the experiment. The actions are inferred through both the think-aloud narration and a separate video processing step; for details see Appendix B. This setting is motivated by the possibility that action recognition could bottleneck model understanding of higher-level cognition.

These conditions allow us to examine whether providing additional structured context about the experiment improves inference of the experimenter's goal-directed cognition, as measured by VQA accuracy. The video-only condition most closely reflects the intended setting in which a model simply observes lab activity, without requiring additional interaction from the experimenter.

Evaluation procedure. For each question, models are prompted to select among three possible answers, based on the provided context. Answer ordering is randomized such that the correct answer may appear in any position. We report total accuracy (chance performance is 33.3%). We use default generation settings and generate one answer per question after finding negligible variance on repeated questions in pilot tests. In our analysis, we ask the following research questions: **RQ1:** How accurately can MLLMs answer the kinds of cognitively grounded questions captured in LabThoughts? **RQ2:** How does providing various levels of context about experimental goals and actions affect model accuracy? **RQ3:** How does performance vary across the components of goal-directed cognition captured by **assessment**, **rationale**, and **intent** statements?

5 Results

Overall accuracy (RQ1). Table 3 reports accuracy across all questions. Gemini 3.1 Pro achieves the highest overall accuracy at 42.5%, followed by Qwen3-VL-32B at 41.4%, while Gemini 3.1 Flash-Lite performs below chance at 23.6%. Performance varies substantially across question types:

[5] https://cloud.google.com/vertex-ai/generative-ai/docs

Model	Goal-Directed Cognition			Temporal		Overall
	Intent	Rationale	Assessment	Intent	Assessment	
Chance	33.3	33.3	33.3	33.3	33.3	33.3
Gemini 3.1 Pro	**60.0** $_{+26.7}$	**29.1** $_{-4.2}$	**43.1** $_{+9.8}$	**49.7** $_{+16.4}$	**39.7** $_{+6.4}$	**42.5** $_{+9.2}$
Gemini 3.1 Flash-Lite	23.3 $_{-10.0}$	16.5 $_{-16.8}$	25.5 $_{-7.8}$	26.2 $_{-7.1}$	29.3 $_{-4.0}$	23.6 $_{-9.7}$
Qwen3-VL-32B	43.3 $_{+10.0}$	26.0 $_{-7.3}$	37.3 $_{+4.0}$	51.9 $_{+18.6}$	44.8 $_{+11.5}$	41.4 $_{+8.1}$

Table 3: **Overall accuracy (RQ1).** Accuracy (%) by model and question type. Temporal questions ask when the experimenter prepared for a later subgoal (**Intent**, **Rationale**) or made an **Assessment** of the experiment state. Subscripts are deviations from chance (33.3%); column best in **bold**.

Context	All questions	
	Acc. (%)	Tokens
Video only	42.4	71,704
+ Goal (text)	42.5	72,193
+ Goal + Actions (text)	**61.3**	76,654

Table 4: **Effect of context (RQ2).** Accuracy and mean reasoning tokens by context. Text descriptions of the actions improves accuracy (+18.9 points). Gemini pro results only; rest in Appendix A.

Context	Goal-Directed Cognition		
	Assessment	Rationale	Intent
Video only	41.2	28.3	36.7
+ Goal (text)	43.1	29.1	60.0
+ Goal + Actions (text)	**72.5**	**71.7**	**80.0**

Table 5: **Accuracy by cognition type (RQ3).** Accuracy (%) on questions grounded in **assessment**, **rationale**, and **intent** statements. Models do best on the lowest-level concept (intent) in every setting. Results shown for Gemini Pro, remainder in Appendix A

Gemini 3.1 Pro performs best on goal-directed intent questions (60.0%), while Qwen3-VL-32B achieves its highest accuracy on temporal intent questions (51.9%).

Effect of additional context (RQ2). Table 4 shows performance across the three context conditions for Gemini 3.1 Pro. Results for other models can be found in Appendix A. Adding the experiment goal alone has essentially no effect on overall accuracy relative to video-only input (42.5% vs. 42.4%). In contrast, providing both the goal and explicit action descriptions increases accuracy to 61.3%, an improvement of 18.9 points over video only.

Accuracy across cognition types (RQ3). Performance differs across **assessment**, **rationale**, and **intent** questions. In the video-only condition, assessment questions are answered with 41.2% accuracy, intent questions with 36.7%, and rationale questions with 28.3%, below chance. With goal and action context provided as structured text, performance increases substantially across all three categories, reaching 72.5% for assessment, 71.7% for rationale, and 80.0% for intent. See Table 5.

Failure modes. We examine the questions that Gemini 3.1 Pro answers incorrectly. We find two recurring failure modes, as illustrated in Figure 6. First, the model often incorrectly chooses an explanation that is locally visually plausible but fails to contextualize it within the broader experimental action sequence, which is needed to infer the underlying intent or assessment. Second, for temporal-localization questions, the model may identify a moment that is procedurally or visually similar to the target event, but does not share the action dependency referenced by the question. These errors suggest that failures may arise from shallow integration of evidence over the course of the experiment, limiting the ability to reason about high-level cognitive states.

6 Discussion

Goal-directed cognition is hard to recover from video. Initial evaluations on the VQA constructed from LabThoughts indicate that goal-directed cognition is not easily recoverable from laboratory video. Further, model accuracy varies substantially across the forms of goal-directed cognition represented in LabThoughts. Specifically, questions grounded in concrete **intentions** are generally answered more accurately than questions grounded in the **rationale** behind those intentions. These

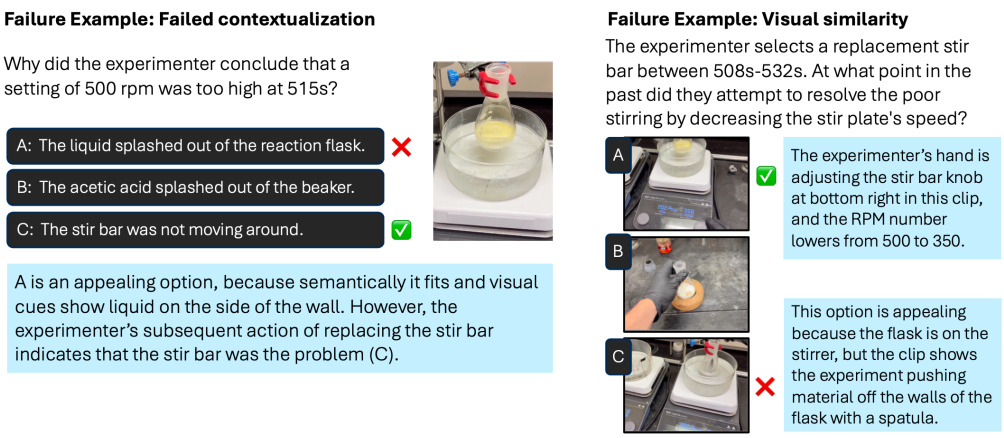


Figure 6: Two common failure modes for Gemini 3.1 Pro. Correct answer denoted by ✓; Gemini's answer denoted by ✗. Elaborations on the failed questions are boxed.

two question types also have different relationships to observable behavior: concrete intentions are often tied directly to a subsequent action in the video, whereas the rationale behind the intent could depend on past experiences, scientific knowledge, or subgoals that are not visually explicit.

Action recognition as a potential bottleneck. We also find that including explicit action recognition scaffolding considerably improves accuracy. Providing a textual description of the experiment goal alone does not improve performance relative to video-only input, whereas additionally providing descriptions of the actions results in a substantial increase in accuracy. Because these actions are in principle recoverable from the video, this result is consistent with action recognition as a bottleneck in reasoning about higher-level goal-directed activity.

Implications for laboratory assistance. Goal-directed understanding may be useful in different ways depending on the laboratory setting. During training, an assistant that understands goals or detects a mismatch between intended and observed activity could potentially provide timely guidance. For experienced researchers, retrospective support may be preferable to proactive intervention. A system that preserves context about why particular decisions were made and how the perceived experimental state evolved could support later troubleshooting of unexpected results or reproducibility failures, as well as brainstorm e.g. about protocol improvements. Such forms of assistance have the potential to accelerate and augment progress in scientific research.

7 Conclusion

We introduce LabThoughts, a dataset designed to study whether MLLMs can understand the goal-directed cognition behind scientific lab work. Unlike prior laboratory datasets that primarily contain actions or prescribed procedures, LabThoughts captures naturally occurring research activity in addition to think-aloud narration and post-experiment interviews. This gives access to researchers' assessments, intentions, and the rationales. We use these statements from the experimenters to construct 574 VQA questions assessing state-of-the-art model insight into experimenter's thought process. The results suggest that MLLMs recover some components of goal-directed cognition more readily than others. They perform comparatively well when identifying concrete intentions, but have lower accuracy on identifying the reasoning behind intentions and the assessments of the evolving experimental state. Providing explicit text summaries of the actions in the video considerably improves accuracy, suggesting that failures in recognizing or integrating low-level activity contribute to low accuracy on the higher-level cognition questions. LabThoughts goes beyond recognizing what a scientist is doing and instead considers how that activity fits into an evolving plan toward a scientific goal. Developing models that can maintain and reason over this goal-directed context will be important for AI systems intended to provide context-sensitive assistance during real scientific work.

AI Disclosure Statement

In this work, we used generative AI tools to assist with implementing the VQA construction and evaluation pipeline. We did not use generative AI tools to develop theoretical or conceptual frameworks, propose or refine hypotheses, design or provide feedback on research methodology or experiments, clean or reformat the dataset, support qualitative or thematic data analysis, or interpret results. Tasks related to mathematical claims, proving mathematical claims, writing proofs, translation, and synthetic dataset generation are not applicable to this work. Additionally, we used generative AI tools to edit text for clarity and readability, assist with software code, identify potential gaps in the literature review, format tables and LaTeX elements, propose keywords for the paper, and compose this AI disclosure statement. All AI-assisted code and pipeline components were reviewed by the authors, and all AI-assisted textual outputs were reviewed and revised as necessary. We take responsibility for the final content of this work, including text, claims, code, and other artifacts produced with the aid of generative AI.

Ethics Statement

The protocol to collect the dataset involves human participants. The protocol was developed with guidance from the IRB and approved by the IRB. Care was taken to de-identify participant data, and precautionary measures such as releasing transcriptions of spoken words in lieu of audio are planned for the dataset release.

Reproducibility Statement

The code to reproduce our VQA pipeline, in addition to the VQA questions, results, and analysis is available at https://anonymous.4open.science/r/cognitive-vqa-8810. The dataset (videos, transcripts, labels) will be released upon acceptance.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.

João Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4724–4733, 2017. doi: 10.1109/CVPR.2017.502.

Le Cong, David Smerkous, Xiaotong Wang, Di Yin, Zaixi Zhang, Ruofan Jin, Yinkai Wang, Michal Gerasimiuk, Ravi K Dinesh, Alex Smerkous, et al. Labos: The ai-xr co-scientist that sees and works with humans. *arXiv preprint arXiv:2510.14861*, 2025.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The dataset. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss (eds.), *Computer Vision – ECCV 2018*, pp. 753–771, Cham, 2018. Springer International Publishing. ISBN 978-3-030-01225-0.

Adele Diamond. Executive functions. *Annual review of psychology*, 64(1):135–168, 2013.

K. Anders Ericsson and Herbert A. Simon. *Protocol Analysis: Verbal Reports as Data*. MIT Press, Cambridge, MA, revised edition, 1993. doi: 10.7551/mitpress/5657.001.0001.

Gianmarco Gabrieli, Irina Espejo Morales, Dimitrios Christofidellis, Mara Graziani, Andrea Giovannini, Federico Zipoli, Amol Thakkar, Antonio Foncubierta, Matteo Manica, and Patrick W. Ruch. Activity recognition in scientific experimentation using multimodal visual encoding. *Digital Discovery*, 4(2):393–402, 2024. doi: https://doi.org/10.1039/d4dd00287c. URL https://www.sciencedirect.com/science/article/pii/S2635098X24002353.

Google DeepMind. Gemini 3.1 flash-lite model card. https://deepmind.google/models/model-cards/gemini-3-1-flash-lite/, March 2026a. Published March 3, 2026.

Google DeepMind. Gemini 3.1 pro model card. https://deepmind.google/models/model-cards/gemini-3-1-pro/, February 2026b. Published February 19, 2026.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Abrham Gebreselasie, Cristina González, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jáchym Kolář, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Ziwei Zhao, Yunyi Zhu, Pablo Arbeláez, David Crandall, Dima Damen, Giovanni Maria Farinella, Christian Fuegen, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18995–19012, June 2022.

Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, Eugene Byrne, Zach Chavis, Joya Chen, Feng Cheng, Fu-Jen Chu, Sean Crane, Avijit Dasgupta, Jing Dong, Maria Escobar, Cristhian Forigua, Abrham Gebreselasie, Sanjay Haresh, Jing Huang, Md Mohaiminul Islam, Suyog Jain, Rawal Khirodkar, Devansh Kukreja, Kevin J Liang, Jia-Wei Liu, Sagnik Majumder, Yongsen Mao, Miguel Martin, Effrosyni Mavroudi, Tushar Nagarajan, Francesco Ragusa, Santhosh Kumar Ramakrishnan, Luigi Seminara, Arjun Somayazulu, Yale Song, Shan Su, Zihui Xue, Edward Zhang, Jinxu Zhang, Angela Castillo, Changan Chen, Xinzhu Fu, Ryosuke Furuta, Cristina Gonzalez, Prince Gupta, Jiabo Hu, Yifei Huang, Yiming Huang, Weslie Khoo, Anush Kumar, Robert Kuo, Sach Lakhavani, Miao Liu, Mi Luo, Zhengyi Luo, Brighid Meredith, Austin Miller, Oluwatumininu Oguntola, Xiaqing Pan, Penny Peng, Shraman Pramanick, Merey Ramazanova, Fiona Ryan, Wei Shan, Kiran Somasundaram, Chenan Song, Audrey Southerland, Masatoshi Tateno, Huiyu Wang, Yuchen Wang, Takuma Yagi, Mingfei Yan, Xitong Yang, Zecheng Yu, Shengxin Cindy Zha, Chen Zhao, Ziwei Zhao, Zhifan Zhu, Jeff Zhuo, Pablo Arbelaez, Gedas Bertasius, Dima Damen, Jakob Engel, Giovanni Maria Farinella, Antonino Furnari, Bernard Ghanem, Judy Hoffman, C.V. Jawahar, Richard Newcombe, Hyun Soo Park, James M. Rehg, Yoichi Sato, Manolis Savva, Jianbo Shi, Mike Zheng Shou, and Michael Wray. Ego-exo4d: Understanding skilled human activity from first- and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19383–19400, June 2024.

Yifei Huang, Guo Chen, Jilan Xu, Mingfang Zhang, Lijin Yang, Baoqi Pei, Hongjie Zhang, Lu Dong, Yali Wang, Limin Wang, and Yu Qiao. Egoexolearn: A dataset for bridging asynchronous ego- and exo-centric view of procedural activities in real world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22072–22086, 2024.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, SOSP '23, pp. 611–626, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400702297. doi: 10.1145/3600006.3613165.

John E. Laird. *The Soar Cognitive Architecture*. The MIT Press, 04 2012. doi: 10.7551/mitpress/7688.001.0001. URL https://cir.nii.ac.jp/crid/1362262944887093248.

Chenyi Li, Guande Wu, Gromit Yeuk-Yin Chan, Dishita G. Turakhia, Sonia Castelo Quispe, Dong Li, Leslie Welch, Cláudio T. Silva, and Jing Qian. Satori: Towards proactive ar assistant with belief-desire-intention user modeling. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pp. 1–24, New York, NY, USA, 2025a. Association for Computing Machinery. doi: 10.1145/3706598.3714188.

Yuxuan Li, Vijay Veerabadran, Michael L. Iuzzolino, Brett D. Roads, Asli Celikyilmaz, and Karl Ridgeway. Egotom: Benchmarking theory of mind reasoning from egocentric videos. *arXiv preprint arXiv:2503.22152*, abs/2503.22152, 2025b. URL https://arxiv.org/abs/2503.22152.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 34892–34916. Curran Associates, Inc., 2023. doi: 10.52202/075280-1516. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/6dcf277ea32ce3288914faf369fe6de0-Paper-Conference.pdf.

Lingni Ma, Yuting Ye, Fangzhou Hong, Vladimir Guzov, Yifeng Jiang, Rowan Postyeni, Luis Pesqueira, Alexander Gamino, Vijay Baiyya, Hyo Jin Kim, Kevin Bailey, David S. Fosas, C. Karen Liu, Ziwei Liu, Jakob Engel, Renzo De Nardi, and Richard Newcombe. Nymeria: A massive collection of multimodal egocentric daily motion in the wild. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision – ECCV 2024*, pp. 445–465, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-72691-0.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 46212–46244. Curran Associates, Inc., 2023. doi: 10.52202/075280-2004. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/90ce332aff156b910b002ce4e6880dec-Paper-Datasets_and_Benchmarks.pdf.

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. HowTo100M: learning a text-video embedding by watching hundred million narrated video clips. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2630–2640, 2019. doi: 10.1109/ICCV.2019.00272.

Toby Perrett, Ahmad Darkhalil, Saptarshi Sinha, Omar Emara, Sam Pollard, Kranti Kumar Parida, Kaiting Liu, Prajwal Gatti, Siddhant Bansal, Kevin Flanagan, et al. Hd-epic: A highly-detailed egocentric video dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23901–23913, 2025.

Herbert A Simon and Allen Newell. Human problem solving: The state of the theory in 1970. *American psychologist*, 26(2):145, 1971.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *CoRR*, abs/1212.0402, 2012. URL http://arxiv.org/abs/1212.0402.

Vijay Veerabadran, Fanyi Xiao, Nitin Kamra, Pedro Matias, Joy Chen, Caley Drooff, Brett Roads, Riley J Williams, Ethan Henderson, Xuanyi Zhao, Kevin Carlberg, Joseph Tighe, and Karl Ridgeway. Benchmarking egocentric multimodal goal inference for assistive wearable agents. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference. Curran Associates, Inc., 2025. doi: 10.52202/085713-0846. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/23ab960082db936f874b171822e0d097-Paper-Datasets_and_Benchmarks_Track.pdf.

Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, Ping Luo, Ziwei Liu, Yali Wang, Limin Wang, and Yu Qiao. Internvid: A large-scale video-text dataset for multimodal understanding and generation. In *The Twelfth*

International Conference on Learning Representations, 2024. URL https://openreview.net/forum?id=MLBdiWu4Fw.

Yicheng Xu, Yue Wu, Jiashuo Yu, Ziang Yan, Tianxiang Jiang, Yinan He, Qingsong Zhao, Kai Chen, Yu Qiao, Limin Wang, Manabu Okumura, and Yi Wang. Expvid: A benchmark for experiment video understanding & reasoning. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 83310–83348, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/86833e97827ff52d297a9fbde5842a81-Paper-Conference.pdf.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025a. URL https://arxiv.org/abs/2505.09388.

Jingkang Yang, Shuai Liu, Hongming Guo, Yuhao Dong, Xiamengwei Zhang, Sicheng Zhang, Pengyun Wang, Zitang Zhou, Binzhu Xie, Ziyue Wang, Bei Ouyang, Zhengyu Lin, Marco Cominelli, Zhongang Cai, Bo Li, Yuanhan Zhang, Peiyuan Zhang, Fangzhou Hong, Joerg Widmer, Francesco Gringoli, Lei Yang, and Ziwei Liu. Egolife: Towards egocentric life assistant. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28885–28900, June 2025b.

Context	All questions	
	Acc. (%)	Tokens
Video only	39.2	0
+ Goal (text)	41.4	0
+ Goal + Actions (text)	**57.9**	0

Table 6: **Effect of context (RQ2).** Qwen 3 VL 32B accuracy and average reasoning tokens by context.

Context	Goal-Directed Cognition		
	Assessment	Rationale	Intent
Video only	29.4	29.1	33.3
+ Goal (text)	37.3	26.0	43.3
+ Goal + Actions (text)	**76.5**	**62.2**	**83.3**

Table 7: **Accuracy by cognition type (RQ3).** Qwen 3 VL 32B accuracy (%) on questions grounded in **assessment**, **rationale**, and **intent** statements.

Context	All questions	
	Acc. (%)	Tokens
Video only	26.5	66,465
+ Goal (text)	23.6	68,139
+ Goal + Actions (text)	**38.3**	70,511

Table 8: **Effect of context (RQ2).** Gemini 3.1 Flash-Lite accuracy and average reasoning tokens by context.

Context	Goal-Directed Cognition		
	Assessment	Rationale	Intent
Video only	27.5	17.3	40.0
+ Goal (text)	25.5	16.5	23.3
+ Goal + Actions (text)	**54.9**	**37.0**	**63.3**

Table 9: **Accuracy by cognition type (RQ3).** Gemini 3.1 Flash-Lite accuracy (%) on questions grounded in **assessment**, **rationale**, and **intent** statements.

A Additional Results

We present additional results on settings and models that were not included in the main text due to space constraints below.

A.1 Effect of context

Results for Gemini 3.1 Flash Lite is in Table 8 and Qwen 3 VL 32B in Table 6.

A.2 Accuracy by cognition type

Results for Gemini 3.1 Flash Lite is in Table 9 and Qwen 3 VL 32B in Table 7.

A.3 Accuracy by experiment

Accuracies across each of the individual experiments are in Table 10.

B Action Scaffolding

We produce a list of actions using the transcript of the experimenter's narration during the experiment, coupled with a hierarchical video understanding pipeline that performs clip segmentation and writes summaries of the actions in those clips.

Concretely, these actions are a flat, list of `scientist_action` entries, with a `start`, an `end`, and a description of what the experimenter is doing; and `state` entries, carrying a `subject` and a `state`, which record an observable condition of the scene rather than an act and may share a span with an action entry. This file, not the automatically inferred timeline, is what later stages of the pipeline treats as the record of what occurred at a given time.

It is produced by reconciling two independently unreliable sources: an automatically inferred timeline of actions produced by a hierarchical video understanding pipeline (Appendix B.1) that produced segmented clips and associated descriptions, and a transcript of the experimenter's narration during the experiment. The transcript establishes *what* is happening and supplies the spoken quantities, but speakers routinely narrate an action before performing it. Conversely, the inferred action description is vulnerable to hallucinated activities and objects.

	Duration	# Questions	Gemini 3.1 Pro	Gemini 3.1 Flash-Lite	Qwen3-VL-32B
			Acc.	Acc.	Acc.
Nitration	43:13	187	42%	22%	36%
Reduction	1:11:45	344	42%	23%	36%
Imine	38:43	226	43%	22%	43%
Bioassay	1:28:29	231	42%	29%	49%
Surfactant	1:12:41	148	43%	19%	50%

Table 10: Accuracy across sessions and models using the video + goal setting.

The reconciliation is therefore performed by a language model (`claude-opus-5`) given both sources, a prior session's output as a schema reference, and the session's stated goal.

Ground-Truth Action Reconciliation

```
You need to create a list of ground truth actions.

The idea is that we know some things about the actions from the
transcript, and some things from inferred_states.json. But
inferred_states.json is guessed by an MLLM, and in the transcript the
person might explain an action before they actually do it. So using
this information, do your best to create a timeline in
ground_truth_actions.

See ../annotations/ground_truth_actions.json for the exact schema/
format to match (a flat list of entries, each either {"type": "
scientist_action", "start": <float seconds>, "end": <float seconds>, "
action": "<description>"} or {"type": "state", "start": <float seconds
>, "end": <float seconds>, "subject": "<what the state describes>", "
state": "<the state itself>"}).

Write the result to .../annotation-videos/ground_truth_actions.json.

When done, report back: how many entries you wrote, what time range
they cover, and flag any real defects, contradictions between the two
sources, or judgment calls you weren't fully sure about, rather than
silently smoothing them over.
```

The resolution strategy the model arrived at was consistent across sessions, and amounts to assigning each source authority over what it can actually observe. Spans are anchored to the clip in which an act is visibly underway, even when the transcript announces it earlier; the transcript overrides the inferred timeline on the identity of apparatus and on quantities and units, with stoichiometric arithmetic used to adjudicate between competing readings of a garbled reagent name; the inferred timeline supplies detail that is visible but never narrated, which is the origin of most `state` entries.

B.1 Automatic State Inference

The automatically inferred state timeline is one of the two inputs to the ground-truth action timeline described in Appendix B. It is produced by two segmentation prompts that split a video into activity segments, and one captioning call per resulting segment. The model used is Gemini 2.5 flash.

B.1.1 Activity Segmentation — Split

Frames are extracted at 1 fps and processed in consecutive chunks of ten. Each chunk is sent with the prompt below, which asks the model to describe the chunk, judge whether it contains one activity or several, and return the 1-indexed frame indices at which activities change. Only the frame list is parsed; the description is discarded.

Activity Segmentation — Split

```
## Context
You will be given a few continuous screenshots of the video
corresponding to approximately 10 seconds of video duration, and
provide detailed, faithful, and accurate analysis of this video
segment. The objective of this analysis is to group the video into
short segments based on the activities for the sake of understanding
the intentions behind the actions.

## Instructions
To perform the analysis of decomposing a video into separate activity
segments, let's do it step by step.
1. Based on the provided frames of this video segment, please describe
 the contents of the video
segment briefly and accurately. You should cover each action and event
 in the clip. The description should be detailed, faithful, and
accurate. It should come with a header: "[1. Description]:".
2. Based on your description, please answer the following question: "
Is this video segment a single activity or combination of multiple
activities?". The definition of an activity is a segment in which the
presenter is working toward a single, nameable unit of progress that a
 viewer would naturally describe with one verb phrase and seek to
independently. It can be an action, for example wiping up a spill. (
However, actions that seem *unrelated to the main task*, e.g. taking a
 sip of a drink, can be ignored.) It can also be an intellectual
activity, for example brainstorming solutions or reading the
introduction to a paper. It ends when that unit is complete or
explicitly abandoned, and a different unit begins. Your answer should
come with a header: "[2. Single:]"
3. If the answer to the previous question is "combination of multiple
activities", please provide the index of frame(s) separating the
scenes from the given frame. Your answer should come with a header:
"[3. Frames]:" and in the format of a list of integers. IMPORTANT:
frames are numbered starting from 1 (1-indexed). The first frame is
frame 1, the second is frame 2, and so on. Use 1-indexed frame numbers
 in your answer. Do NOT include frame 1 in the list - only report
boundaries at frame 2 or later, since frame 1 is always the start of
this clip and is handled separately.

## Example
Your response should be in the following format:
[1. Description]: This video shows ...
[2. Single: yes/no]: No.
[3. Frames]: [5, 9]
Please pay special attention to:
- The precise localization of the frames is very important for
downstream tasks.
- Frames are 1-indexed: the first frame you are given is frame 1, not
frame 0.
- The summarization at the scene level should be consistent with the
frames you provided. For instance, the number of scenes should be one
more than the number of frames in the list. If you provide 0 frames
since the images display a consistent scene, you will give 1 summary;
If you provide 1 frame, there should be 2 summaries; if you provide 2
frames, there should be 3 summaries, etc. Now you will be presented
the video frames, please perform the analysis carefully.
```

B.1.2 Activity Segmentation — Merge

Splitting within fixed ten-frame chunks makes every chunk boundary a provisional activity boundary. For each junction, the last frame of the preceding chunk and the first frame of the following chunk are sent with the prompt below; an affirmative answer deletes the boundary, allowing a single activity

to span multiple chunks. Both prompts share one definition of an activity, so that the split and merge decisions are made against the same criterion. Model: `gemini-2.5-flash`.

Activity Segmentation — Merge

```
You are going to help with determining if a short video segment is a
consistent activity. You will be given a few continuous screenshots of
 the video clip, along with the accompanying audio, and provide
detailed, faithful, and accurate analysis of this video segment. Your
objective is simple: if *the video clip starting from the second frame
* is a consistent activity with *the first frame*. Answer with "yes"
or "no". The definition of an activity is a segment in which the
presenter is working toward a single, nameable unit of progress that a
 viewer would naturally describe with one verb phrase and seek to
independently. It can be an action, for example wiping up a spill. (
However, actions that seem *unrelated to the main task*, e.g. taking a
 sip of a drink, can be ignored.) It can also be an intellectual
activity, for example brainstorming solutions or reading the
introduction to a paper. It ends when that unit is complete or
explicitly abandoned, and a different unit begins.  Now you will be
presented the video frames, please perform the analysis carefully.
```

B.1.3 Per-Segment Captioning

The surviving segment boundaries are converted from frame indices to video time and the recording is cut into one clip per activity. Each clip is then sent on its own — video only, with no transcript, no goal description, and no context from neighbouring clips — with the prompt below. The response is constrained to a two-field JSON schema, and the resulting records form `inferred_states.json`. Model: `gemini-2.5-flash-lite`, temperature 0.2.

Per-Segment Captioning

```
You are captioning one clip from a longer recording.

Return ONLY a JSON object with two fields:
- clip_summary: a factual, grounded description of this clip - what is
 on screen, what happens,
  what is said. Describe in detail, as though captioning. It is a "
summary" only because it is
  less detailed than the full clip, not because it should be vague.
- current_action: as specific as the evidence allows (e.g. "Rephrasing
 the opening paragraph of
  the Methods section" rather than "Editing a document"). Only assert
what is directly
  observable.
```

C Prompts

The prompts used in the VQA pipeline are listed in §C.1 and §C.2, while prompts used in evaluation are given in §C.3.

C.1 Question Generation

C.1.1 Goal-Directed Cognition Questions

Given a transcript comment and its surrounding context, writes an open-ended question that tests why the experimenter is doing something, not recall of protocol facts.

Goal-Directed Cognition: All annotated lines

```
    You are writing a multiple choice test for advanced, graduate-
level scientists.
    You will be given a scientific experiment and some context.
    You will also be given the experimenter's real comments about a
certain part of the experiment.

    Your task is to create a multiple choice question based on that
comment and the surrounding context. The question should test the
student's ability to anticipate and understand what a real
experimenter is thinking during this specific experimenter. The
students will not have access to the experimenters' comments, only the
 video of the experiment itself. Therefore, the question should not
reference the comments at all, only the video. The goal is to test
whether the test takers have developed a seasoned scientific
understanding of the experiment that matches the experimenters'.

    Do not include any context at all in the question. For instance,
it is bad to include in the question a phrase like "Given that this is
 a nitration."

    You will also be given background reference material (a lab
protocol, or other written or spoken
    description of the experiment's goal) in addition to the video.
This material often contains specific
    facts - reagent names, volumes, step order, what happens next. Do
NOT write a question that merely asks
    the test-taker to recall or predict one of these facts (e.g. "What
 volume are they using?", "What reagent
    comes next?", "What material is X, and what is inside it?", "What
do these numbers represent?") - that
    tests memorization of the reference material, not scientific
understanding. Every question must still
    test insight into the experimenter's thinking: WHY they are doing
what they're doing, what risk or
    precaution they're managing, or what they anticipate will happen
next and why - never a bare fact.

    Bad example (tests recall of the reference material, not reasoning
):
    Comment: "This bottle has silicon nanoparticles in it, so I'm
wrapping the cap with Parafilm just to be safe."
    Question: "What material is the experimenter wrapping around the
bottle, and what is likely inside?"
    Why this is bad: it only asks the test-taker to name the material
and contents. It doesn't test whether
    they understand WHY wrapping it matters - the precaution against
the specific risk those nanoparticles pose.
    Better: "Why is the experimenter taking the precaution of wrapping
 this bottle's cap?"

    For example:
    Comment: "So the reaction hasn't started yet though, because we
still need to add our nitric acid."
    Question:"Why hasn't the reaction started yet?"
    Question: "Has the reaction started? Why or why not?"
    Rationale: The experimenter notes that the reaction hasn't started
. Test whether the student knows this. The experimenter notes that the
 reaction hasn't started because the acid hasn't been added yet. Test
whether the student knows this.

    Comment: "Usually with the addition of the tin and the HCL, things
 will start to loosen up and dissolve a bit"
```

```
    Question: "What is the likely effect of the addition of tin and
HCL on the solution?"
    Rationale: The experimenter knows that adding tin and HCL will
loosen up the solution. Test whether the student also knows this.
    Question: "Why is the stirring likely to improve in the future?"
    Rationale: The context of the experimenter's comments makes it
clear that this is in reference to anticipating better stirring in the
 future. Test whether the student knows this.

    Comment: "And as we're doing that, we should notice that the
solution should kind of go to a more transparent yellowy color."
    Question: "What kind of color change should we expect to see in
the solution?"
    Rationale: The experimenter is anticipating that in the future the
 solution will be more transparent and yellowy. Test whether the
student knows this.

    Return ONLY a JSON object in this exact format:
{
  "question_answer_pairs": {
    "question": "\<The multiple choice question that tests
understanding of the experimenter's internal thoughts and comment\>",
    "answer": "\<The correct answer. The incorrect options will be
created later.\>",
    "rationale": "\<The part of the experimenter's comments that the
question is testing.\>"
  }

    Context (The experimenter's stated goal for the experiment): {
goal_block}
    Context (Summary of what has happened in the experiment up to this
 point): {context_block}
    Context (Video at this point): {window}
    Comment from the experimenter at this point in the experimenter: {
transcript_text}
```

C.1.2 Temporal Localization: Only annotations labeled intents and rationales

Given a comment in which the experimenter is preparing for a future outcome, identifies the later point in the video that outcome corresponds to and creates a question asking the test taker to locate the earlier point where the experimenter was preparing for it.

Temporal Dependency

```
    You are preparing to create an advanced quiz that tests
understanding of dependencies in a scientific experiment.

    You will be given a summary of the whole experiment, along with
the experimenter's thoughts at a certain point in the experiment.
These thoughts will be related to something that will happen in the
future -- for instance, how the experimenter is currently preparing
for something to happen to in the future, or else how the experimenter
 is currently preparing to ensure that something negative does not
happen in the future.

    Your task is to find which point(s) in the experiment corresponds
to that future scenario that the exerimenter is currently anticipating
.

    Then, you will create a question for a student. Imagine the
student is watching a video of the experiment, and has reached the
future point you identified. The question should ask the student to
identify at what point in the past the experimenter was preparing for
```

```
this outcome (or lack of negative outcome). Note that the student will
 not have access to the experimenter's comments, only the video of
their actions. Your question should assess how aligned the student is
to the experiment's planning.

    Do not include any context at all in the question. For instance,
it is bad to include in the question a phrase like "Given that this is
 a nitration."

    Return ONLY a JSON object in this exact format:
{
  "dependency_question": {
    "future_event": "\<Timestamp of the future scenario you identified
\>",
    "question": "\<Asking about when the experimenter acted to prepare
 for the future event\>",
    "answer": "\<Timestamp of the point in the video that answers the
question\>",
  }

    Context (The experimenter's stated goal for the experiment): {
goal_block}
    Description of the whole experiment: {context_block}

    Experimenter's thoughts at a certain point in the experiment: {
transcript_text}
```

C.1.3 TEMPORAL LOCALIZATION: ONLY LINES LABELED ASSESSMENT

Locates the discrete, locally distinguishable visual trigger for the assessment expressed by the experimenter and asks the test taker to identify when it occurred, or declines if no suitable trigger exists.

Stage 3 — Assessment

```
You are preparing to create an advanced quiz that tests PhD candidates
 understanding of an ongoing scientific experiment.

    You have the COMPLETE video of a real experiment, along with the
experimenter's thoughts at one particular point in it. The comment in
question is SPOKEN from {reference_start:.0f}s to {reference_end:.0f}s
 in that video; locate it yourself and study what is on screen there,
and in the minute or so before and after -- the cue being discussed
often sits outside the span of the speech itself. These thoughts will
be related to an assessment of the current experimental state -- for
instance, whether the reaction is proceeding normally, or whether the
experimenter is noticing a possible risk that should be addressed.

    Judge both properties below against the VIDEO, not against the
timestamp list. You can see the
    whole experiment, so check the neighbouring footage directly
rather than reasoning about what
    the list implies. The test taker will see this same video at this
same quality: if you cannot
    make out a cue yourself, neither can they, and the item is
unusable.

    Your task is to create a question that asks the test taker to
identify the timestamp at which the experimenter made this assessment.
 The test takers will not have access to the experimenter's thoughts;
they will only see the video of the experiment itself. We want to test
 whether they have developed a seasoned scientific understanding such
that their reading of the experiment matches the experimenter's.
```

```
    A good item for this test requires TWO properties. Check both
before writing anything.

    (1) A DISCRETE VISUAL TRIGGER. There must be something concretely
observable on screen -- a
        change in the apparatus, the mixture, an instrument readout,
or the experimenter's own
        action -- that a careful viewer could point to as the reason
the assessment was made.

        The timestamps you are given mark when the experimenter SPOKE,
 which is often NOT when
        the cue was on screen. It can fall either side. Sometimes they
 narrate something they
        have already seen, so the trigger comes first -- by a few
seconds, occasionally by half a
        minute -- and the camera has moved on by the time the words
are said. Other times they
        announce first and then show, tilting the flask toward the
lens or panning to a readout
        just after speaking. So search roughly the 45 seconds before
and the 30 seconds after the
        comment, and NEVER conclude a cue is absent merely because it
is off screen during the
        speech itself. Report the moment you actually found it in
trigger_timestamp.

        The trigger also only has to be PART of what prompted the
assessment, not the whole of
        it. Experimenters constantly combine what they see with
arithmetic, prior knowledge, or
        memory of earlier steps: reading a level off a graduated vial
and subtracting to get
        what remains is still triggered by something on screen. Do not
 reject an item because
        the conclusion required a calculation or expertise to reach --
 that is what makes it
        worth testing.

        The widened window is for finding THIS comment's cue, not for
finding a better comment.
        Whatever you settle on must be the thing THIS comment is about
: if the remark is about
        the mixture's color, the trigger is that color; if it is about
 a temperature reading, the
        trigger is that readout. A vivid, easily-described event
happening nearby is NOT a
        licence to build the question around it instead. Ask yourself:
 would the experimenter
        have said these exact words because of the thing I picked? If
not, you have the wrong
        trigger, however photogenic it is.

        It follows that when the comment carries no assessment at all
-- logistics ("I need new
        gloves"), narration of one's own action, a statement of what
comes next -- the answer is
        to DECLINE. Do not rescue such a moment by substituting a
nearby event that does carry an
        assessment; that produces a question the source comment cannot
 answer. Declining is the
```

```
        correct, expected outcome here, and it is much better than a
fabricated item.

        "Logistics" here means a remark with NO observable basis --
housekeeping the experimenter
        would have said regardless of what the apparatus was doing. It
 does NOT mean any remark
        that mentions a quantity or a step. "6.3 mils, we probably
added like 1.3, so we probably
        still have about 5 mils left" is read off the graduations of a
 vial the camera is looking
        at; it is an assessment of how far the addition has got, and
it is exactly the kind of
        quantitative read-plus-inference worth testing. Before calling
 something logistics, ask
        whether the experimenter had to LOOK at anything to say it. If
 they did, it is not
        logistics.

        Property (1) fails only if nothing anywhere in that
surrounding window is observable --
        an assessment the experimenter merely thought, anticipated, or
 remembered, with no
        on-screen correlate at all.

    (2) LOCAL DISTINGUISHABILITY. The test taker will choose between
timestamps that are at
        least 100 seconds apart, so the trigger only needs to be
distinguishable from its own
        neighbourhood -- not from the entire experiment. A slow or
continuous change is
        perfectly acceptable: a mixture that yellows steadily over
five minutes, a temperature
        that holds flat for a long stretch, or a solid that stays
visible throughout are all
        fine, so long as this is the point at which the experimenter
would actually remark on
        it. Reject only when a distinctly separate, comparably good
moment sits within roughly
        two minutes and would prompt the same inference just as well.

    If either property fails, DO NOT invent a question. Return the
unsuitable form described
    below instead. It is much better to decline than to produce an
item whose answer cannot be
    defended; many candidate moments in an experiment are legitimately
 unsuitable.

    If both properties hold, write the question. The question must
name the INFERENCE the
    experimenter drew, not the percept they saw. A question that
restates the observation turns
    the task into visual search and tests nothing about scientific
judgment.

    Be careful not to over-apply this. Most assessments in a lab are
phrased AS an observation
    with a verdict attached, and the verdict is the inference -- it
does not have to be spelled
    out as a separate sentence. "This orangier state is common" says
the color is EXPECTED;
    "that's a really good sign" says the reaction is on track; "I'm
not too worried about that"
```

```
    says a visible anomaly is benign. All of those are assessments,
and all are fair game. You
    are also given the couple of transcript lines preceding the
comment: use them. A short line
    like "Getting more yellow" may look like a bare percept alone
while plainly continuing a
    running series of progress checks in context. Reject for this
reason only when the comment
    carries no evaluative content at all, in its context -- a pure
naming of what is on screen.

    BAD  (restates the percept): "At what timestamp does the reaction
mixture visibly become
         more transparent?"
    GOOD (names the inference):  "At what moment does the experimenter
 first have visual
         confirmation that the reaction is proceeding as intended?"

    BAD  (restates the percept): "At what timestamp does the
experimenter put on a fresh pair
         of gloves?"
    GOOD (states the reason behind the observation): "At what point
does the experimenter realize that there may be a risk of chemical
transfer?"

    Worked example of a suitable item:

    Experimenter comment: It means we probably are getting a little
bit of product in there.
    Comment spoken at: 2218s
    trigger_timestamp: 2204s
    visual_evidence: The contents of the flask shift to a lighter
color, visible through the glass. This happens at 2204s, about
fourteen seconds before the experimenter remarks on it -- by 2218s the
 camera has already drifted toward the bench.
    why_not_elsewhere: No other point within a couple of minutes
either side would prompt this inference -- the mixture is a steady
darker color beforehand, and nothing else in that neighbourhood
suggests product is forming. The color continues to deepen later on,
but this is where the experimenter would first call it.
    Question: What moment in the video gives the experimenter the
first evidence that product is forming?
    Answer: 2218s
    Rationale: The experimenter connects the color change to product
formation. Test whether the students can draw the same inference from
the same visual cue.

    Do not include any context at all in the question. For instance,
it is bad to include in the question a phrase like "Given that this is
 a nitration."

    Context (The experimenter's stated goal for the experiment): {
goal_block}

    Segment boundaries for the whole experiment, as bare timestamp
spans. These tell you how the
    experiment is carved up, NOT what happens in each span -- the
descriptions are withheld on
    purpose, because the test taker doesn't get them either. Anything
you claim is observable must
    come from watching the video:
{ground_truth_actions}
```

```
    Experimenter's comment at this point ({reference_start:.0f}s-{
reference_end:.0f}s). The line
    inside \<focus\>...\</focus\> is THE ONE you are assessing, and
the timestamps above are its
    timestamps. Any lines before it are earlier narration, included
only so you can tell what the
    focused line means -- they are NOT what you are assessing. Build
the question around the
    focused line alone. If the focused line carries no assessment,
DECLINE: do not reach back to
    a context line that does, because the answer would then point at
the focused line's
    timestamp while the question described a different moment.
{transcript_text}

    If the moment is suitable, return ONLY a JSON object in this exact
 format:
{
  "assessment_question": {
    "trigger_timestamp": "\<When the cue is actually ON SCREEN, in
seconds. This need not fall inside the spoken comment's span -- say
where you really found it.\>",
    "visual_evidence": "\<What is concretely observable on screen at
that timestamp that triggered the assessment\>",
    "why_not_elsewhere": "\<Why no nearby moment -- roughly two
minutes either side -- fits this question as well as the one you chose
. A gradual change is fine; say why this is the point the experimenter
 would remark on it. Refer to the surrounding footage and the ground
truth actions.\>",
    "question": "\<Asking when the experimenter drew this inference
about the experimental state\>",
    "answer": "\<Timestamp of the point in the video\>",
    "rationale": "\<Explanation of what the question is testing\>"
  }
}

    If the moment fails either property, return ONLY:
{
  "assessment_question": {
    "unsuitable": "\<Which property fails and why\>"
  }
}
```

C.2 Answer (Distractor) Generation

C.2.1 Context Free Distractor

Generates an additional plausible prose distractors from the question, correct answer, and experiment goal alone, with no access to the video or transcript.

Blind (prose answers)

```
You will be given a single quiz question about a scientific chemistry
experiment, along with
a possible answer and the experimenter's stated goal for the
experiment. You do NOT have
access to the video or any other of the experimenter's comments - only
 the question, one
possible answer, and the stated goal of the experiment.

Provide TWO additional plausible answers to the question. They must be
 genuinely different from both one another and from the provided
```

```
answer. No answer should be a rephrasing or partial restatements of
another answer.

CRITICAL FORMATTING RULE: your two answers MUST match the provided
answer's format and length
almost exactly. The provided answer is {correct_answer_word_count}
words long - keep each of
your answers within plus or minus 10 words of that. If the provided
answer is a short phrase (e.g. "to
clear out the funnel"), write short phrases, NOT full sentences. If it
's one or more full
sentences, match that instead. An answer that is noticeably longer,
grammatically different, or more
explanatory than the provided answer is WRONG regardless of its
content, because it makes one answer stick out too much - brevity and
register must match, not just
plausibility.

Return ONLY a JSON object in this exact format:
{
  "answer1": "\<first plausible answer, matching the provided answer's
 format and length\>",
  "answer2": "\<second, genuinely different plausible answer, matching
 the provided answer's format and length\>"
}

Experiment goal: {goal_context}
Question: {question}
Possible answer: {correct_answer}
```

C.2.2 BLIND (TIMESTAMP ANSWERS)

Chooses two plausible incorrect timespans from the listed timestamped actions alone, with no access to the video.

Blind (timestamp answers)

```
You will be given a quiz question about a scientific chemistry
experiment, its correct answer,
and a list of timestamped actions/observations from that experiment (
the "action sheet"). The
answer to the question is one point in time. You do NOT have access to
 the video or the
experimenter's comments - only the question, one possible answer, and
the action sheet.

From the action sheet, choose the TWO timespans that are the most
plausible INCORRECT answers to
the question. They must be two DIFFERENT entries, and neither may be
the correct answer's
timespan. Pick tempting, on-topic times - not obviously unrelated ones
.

Return ONLY a JSON object in this exact format, copying each timespan
exactly as it appears in
the sheet (e.g. "246s-279s"):
{
  "answer1": "\<timespan of the first plausible incorrect entry\>",
  "answer2": "\<timespan of the second, different plausible incorrect
entry\>"
}

Question: {question}
```

```
Possible answer: {correct_answer}

Action sheet:
{action_sheet}
```

C.2.3 GROUNDED (PROSE ANSWERS)

Given the video clip, replaces a generic placeholder distractor with a specific, video-grounded incorrect answer whose core claim conflicts with the correct answer's.

Grounded (prose answers)

```
You are helping write a detailed answer key for a quiz about a
scientific chemistry
experiment. You have been given a video clip of the relevant moment of
 the experiment and
the experimenter's stated goal for it.

Below is a quiz question and a GENERIC answer to it - a vague
placeholder that was written
without watching this particular experiment, and that could apply to
many experiments.
Unlike whoever wrote it, you can see this specific moment.

IMPORTANT: a more specific, more detailed, or more concrete version of
 the generic answer is
ALWAYS WRONG for this task, no matter how accurate it is. Simply
naming the specific
equipment, reagent, or quantity involved in the SAME underlying claim
as the generic answer
does not count as a different answer - it is still the same answer,
just dressed up, and
must not be submitted. Your answer's core claim (the actual reason,
mechanism, or judgment it
gives) must be a DIFFERENT one from the generic answer's core claim -
one that could not both
be true at once. For example, if the generic answer's claim is "yes,
this is desirable", your
answer must not also claim "yes, this is desirable" with extra detail
added; it must argue
for a different verdict or a different underlying reason.

Your job is to invent that different, incorrect claim, and then ground
 IT - not the generic
answer - in specifics from what you can actually observe in the clip
and the stated goal, so
that it reads with the same real, concrete confidence a true answer
would have: naming
particular equipment, materials, quantities, or conditions you see,
but in service of a claim
that is actually false for this moment.

Write ONE such answer. Before finalizing it, check yourself: if
someone believed only your
answer, would they conclude something different from, and incompatible
 with, the generic
answer? If your answer could be true at the same time as the generic
answer, it has failed -
revise it so the two genuinely conflict.

CRITICAL FORMATTING RULE: be specific in CONTENT, not in length. Your
answer MUST match the
```

```
generic answer's format and length almost exactly. The generic answer
is
{correct_answer_word_count} words long - keep your answer within plus
or minus 10 words of
that. If the generic answer is a short phrase (e.g. "to clear out the
funnel"), write a short
phrase, NOT a full sentence. If it is one or more full sentences,
match that instead. An
answer that is noticeably longer, grammatically different, or more
explanatory than the
generic answer is WRONG regardless of its content, because it makes
one answer stick out too
much - brevity and register must match, not just plausibility.

Return ONLY a JSON object in this exact format:
{
  "answer": "\<the specific, incorrect answer, matching the generic
answer's format and length\>"
}

Question: {question}
Generic answer: {correct_answer}
```

C.2.4 GROUNDED (TIMESTAMP ANSWERS)

Given the video and a timespan already known to be wrong, finds the video-grounded runner-up timespan the model would otherwise mistake for the correct answer.

Grounded (timestamp answers)

```
You are helping repair the answer key for a quiz about a scientific
chemistry experiment. You
have been given the video of the experiment and the experimenter's
stated goal for it. Times
in the video and in every timespan below are in the same coordinates:
seconds from the start
of the original recording.

Below is a quiz question whose answer is a single point in time, and
the timespan the answer
key currently gives for it. That timespan has already been checked and
 is WRONG - the moment
the question asks about does NOT happen there. Unlike whoever wrote
the key, you can watch
the video. Your job is to find where in the video this ACTUALLY takes
place.

Watch the video, work out which moment the question is really pointing
 at, and report the ONE
timespan that covers it. Your answer must be copied exactly from the
candidate list below -
these are the experiment's event boundaries (the known-wrong timespan
has already been
removed from the list). Pick the single best one, and pick it from
what you SEE happening at
that time, not from a guess about where in the experiment it ought to
fall.

Return ONLY a JSON object in this exact format, copying the timespan
exactly as it appears in
the list (e.g. "246s-279s"):
{
  "answer": "\<the timespan where this actually takes place\>"
```

```
}

Question: {question}
Timespan the answer key gives (already verified WRONG): {
correct_answer}

Candidate timespans:
{candidate_timespans}
```

C.3 Evaluation

C.3.1 Oracle

The oracle evaluation setting: the judge model answers using both the video and all available text context, including privileged transcript.

Oracle

```
Answer the following multiple-choice question based on the video and
the provided context of the experimenter's comments.

Question: {question}

{choices_block}.

Respond only with the letter of the best answer to the question.
```

C.3.2 Only Video

Video only; no textual context (transcript or ground-truth action log) is provided.

Only Video

```
Answer the following multiple-choice question based ONLY on the video.
 No textual context
about the experimenter's comments or the experiment's ground-truth
action log is provided -
rely solely on what you observe in the video.

Question: {question}

{choices_block}.

Respond only with the letter of the best answer to the question.
```

C.3.3 No context

Neither video nor textual context; the fully blind floor condition.

No context

```
Answer the following multiple-choice question about a chemistry
experiment to the best of your ability. You will NOT be shown a video,
 but you can imagine how the events of the experiment might plausibly
unfold over its roughly {video_duration}.

Question: {question}

{choices_block}.

Respond only with the letter of the best answer to the question.
```